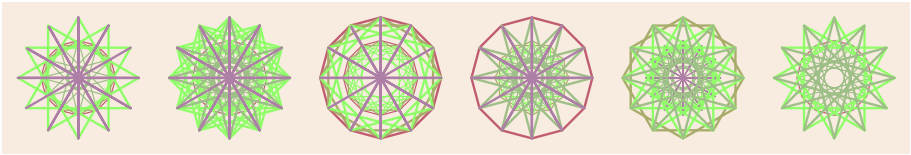
Example

Leon Tabak

08 February 2022



1. What role has Aurélien Géron asked us to take as we read Chapter 2 of *Hands-On Machine Learning*?

2. Chapter 2 of *Hands-On Machine Learning* begins with a list of steps. The author has elaborated this list in Appendix B. The appendix begins on page 755.

   (a) The word "business" appears 3 times in the section titled **Frame the Problem and Look at the Big Picture**. In what ways?

   (b) The words "legal," "authorization," and "sensitive" appear in the section titled **Get the Data.** What advice does Géron give us here?

   (c) In the **Explore the Data** section, the author names some types of noise and some types of probability distributions. Are any of these types already familiar to you? Can you learn what some of the words that are new to you mean?

   (d) In **Prepare the Data**, the author gives several options for dealing with missing values. What are they?

   (e) What does Géron tell us to do in a "quick and dirty" manner in **Shortlist Promising Models**?

(f) The section titled **Fine-Tune the System** mentions "Ensemble methods." Can you guess from the context and text that follows what an ensemble method is?

(g) In **Present Your Solution**, you will see the words "make sure" and "explain why" and "don't forget." Elaborate. What are we supposed to do?

(h) What is the meaning of the word "rot" where it appears in the **Launch!** section?

3. The California Housing Prices dataset is old. How old?

4. The dataset contains information about block groups. What is a block group?

5. A footnote in Chapter 2 mentions Claude Shannon. He is a very interesting person. Find out something about him to share with the class.

   A biography of Shannon appeared a few years ago. It won prizes. Can you find the title?

6. Why are we building a model of housing prices? Where will the output of our model go? Our model will help our clients make what kinds of decisions?

7. We will use *multiple regression*. What does that mean?

8. What are the arguments for batch learning in this case?

9. Here is the definition of Root Mean Square Error:

$$RMSE(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1} m\left(h(\mathbf{x}^{(i)} - y^{(i)}\right)^2}$$

   (a) Which letter represents the whole dataset (minus the labels)?

   (b) Which letter represents on row in the dataset (minus the label)?

   (c) Which letter represents a label? (That's a price in this case.)

   (d) Which letter represents the number of records? (That's the number of rows in out table of data.)

   (e) Which letter represents the function that predicts the price of homes in a district, given a description of that district?

10. RMSE is an example of a way of measuring the distance between two vectors. This kind of measure is also called a "cost function" and a "norm."

    Other names for RMSE are "Euclidean norm" and the ℓ_2 norm.

    RMSE is a good choice for normally distributed data.

    What is an alternative to RMSE that author suggests for cases in which data contains many outliers?